第二次作业**2016-10-17**

本次作业一共5道题目，前4题为计算证明题，最后一题为上机题。

Note：

（1）作业统一以pdf格式提交，命名为**学号_姓名.pdf,**如“201628014628053_吴金文.pdf “。程序源码等打包到**学号_姓名.zip**提交。

（2）上机题需要提交源码，并指出运行环境以及环境依赖以方便查看。源码中建议提供简单注释。

（3）作业时间为2周, 通过选课网站http://sep.ucas.ac.cn/，在对应课程的**课堂作业**栏目下提交。若提交时间有变动，网站上会通知。

---

1. 本题有两小题。

（1）设一维特征空间中的窗函数 $\varphi(u)=\begin{cases}1, & |u|<1/2\\ 0, & \text{otherwise}\end{cases}$，有 n 个样本 x_i, i=1,…,n, 采用宽度为 h_n 的窗函数，请写出概率密度函数 p(x)的 Parzen 窗估计 p_n(x)；

（2）给定一维空间三个样本点{-1,0,2}，请写出概率密度函数 p(x)的最近邻（1-NN）估计并画出概率密度函数曲线图。

2 . Consider data $\mathcal{D}=\{\binom{1}{1},\binom{3}{3},\binom{2}{*}\}$, sampled from a two-dimensional (separable) distribution $p(x_1,x_2)=p(x_1)p(x_2)$, with

$$p(x_1) \sim \begin{cases}\frac{1}{\theta_1}e^{-\theta_1 x_1} & \text{if } x_1\geq 0\\ 0 & \text{otherwise,}\end{cases}$$

and

$$p(x_2) \sim U(0,\theta_2)=\begin{cases}\frac{1}{\theta_2} & \text{if } 0\leq x_2\leq\theta\\ 0 & \text{otherwise.}\end{cases}$$

As usual, $*$ represents a missing feature value.

(a) Start with an initial estimate $\boldsymbol{\theta}^0=\binom{2}{4}$ and analytically calculate $Q(\boldsymbol{\theta},\boldsymbol{\theta}^0)$ — the **E step** in the EM algorithm. Be sure to consider the normalization of your distribution.

(b) Find the $\boldsymbol{\theta}$ that maximizes your $Q(\boldsymbol{\theta},\boldsymbol{\theta}^0)$ — the **M step**.

(c) Plot your data on a two-dimensional graph and indicate the new parameter estimates.

3. 用最大似然法估计类别 ω_i 的先验概率 $P(\omega_i)$。随机、独立地抽取 n 个样本，如果第 k 个样本属于 ω_i， $z_{ik}=1$，否则 $z_{ik}=0$。

(1)写出 $P(z_{i1},\ldots,z_{in}\,|\,P(\omega_i))$ 的表示式。

(2)给出 $P(\omega_i)$ 的最大似然估计。

4. 现有两类二维样本如下：

w1：(-1,0), (-2,0), (-2,1), (-2,-1), (-3,-1), (-2,0.5), (-2,-0.5), (0,0)

w2: (-1,0), (0,0), (1,1), (2,1), (2,-1)

(1) 请分别采用 1 近邻和 3 近邻设计分类器

(2) 可能出现不同类的样本都是某个点的近邻的情形。针对此情形，请采用拒绝分类规则重新设计分类器。

(3) K 近邻算法的优缺点是什么。

5. 实验题：

请用 KNN（k-nearest neighbor）对 MNIST 数据进行分类，比较不同参数下的结果并讨论。请把实验结果和相应讨论写在提交的 pdf 中。

MNIST 数据集：http://yann.lecun.com/exdb/mnist/